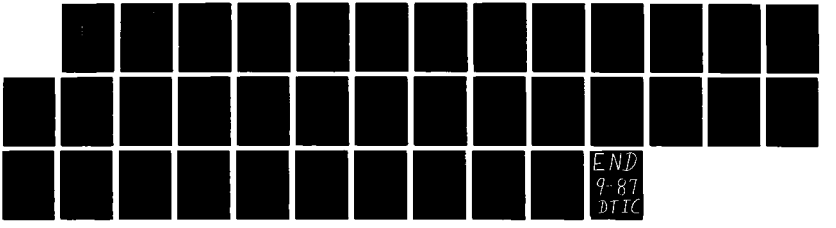
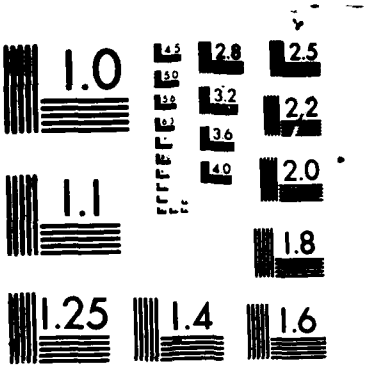
AD-A183 796 THE LEXICON IN TEXT GENERATION(U) UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST S CUMMING OCT 86 ISI/RR-86-168 MDA903-81-C-0335 1/1

UNCLASSIFIED F/G 5/7 NL

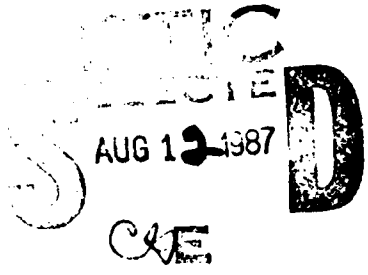


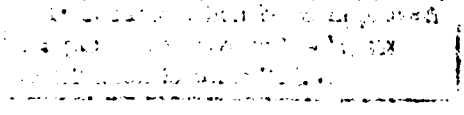


ISI Research Report
ISI/RR-86-168
October 1986

Susanna Cumming

The Lexicon in Text Generation





INFORMATION SCIENCES INSTITUTE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT This document is approved for public release; distribution is unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ISI/RR-86-168	5. MONITORING ORGANIZATION REPORT NUMBER(S) ---
6a. NAME OF PERFORMING ORGANIZATION USC/Information Sciences Institute 6b. OFFICE SYMBOL (*If applicable*)	7a. NAME OF MONITORING ORGANIZATION ---
6c. ADDRESS (*City, State, and ZIP Code*) 4676 Admiralty Way Marina del Rey, CA 90292	7b. ADDRESS (*City, State, and ZIP Code*) ---
8a. NAME OF FUNDING / SPONSORING ORGANIZATION Advanced Research Projects Agency 8b. OFFICE SYMBOL (*If applicable*)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903 81 C 0335
8c. ADDRESS (*City, State, and ZIP Code*) 1400 Wilson Boulevard Arlington, VA 22209	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. --- / PROJECT NO. --- / TASK NO. --- / WORK UNIT ACCESSION NO. ---
11. TITLE (*Include Security Classification*) The Lexicon in Text Generation (Unclassified)	
12. PERSONAL AUTHOR(S) Cumming, Susanna	
13a. TYPE OF REPORT Research Report 13b. TIME COVERED FROM _____ TO _____	14. DATE OF REPORT (*Year, Month, Day*) 1986, October 15. PAGE COUNT 37
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
09	02	

18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*)
artificial intelligence, categories, computational linguistics, dictionary, features, lexical semantics, lexicon, natural language, subcategorization, syntax, text generation

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

(over)

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☑ UNCLASSIFIED/UNLIMITED ☑ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Sheila Coyazo Victor Brown	22b. TELEPHONE (*Include Area Code*) 213-822-1511 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

This report compares several lexicons used in computational text generation systems, with respect to the size of the lexical item, the way coocurrence phenomena are represented, and the way semantic information is included. The lexicons examined can be roughly divided into two principal groups with respect to the size of the item, "phrasal lexicons" and "word-based lexicons". Phrasal lexicons, which are more numerous, have large units (sometimes whole sentences) stored as lexical entries. They often tend to represent syntactic structure within the lexical item, and may also contain variables or slots which can be filled by other items. This type of lexicon generally provides the primary line between semantic and syntactic representation by mapping semantic structures onto syntactic structures. The word-based lexicon, on the other hand, merely inserts words into previously built syntactic structures, using feature specifications to guide the process.

Lexicons also vary with respect to the amount of coocurrence information they contain. Most lexicons represent subcategorizational (argument structure) information, either by means of features or with syntactically labelled slots. They can also have noncompositional multi-word units (idioms) as lexical entries. Some lexicons represent selectional information as well, by means of semantic feature restriction on slots. Collocational information is rarely included.

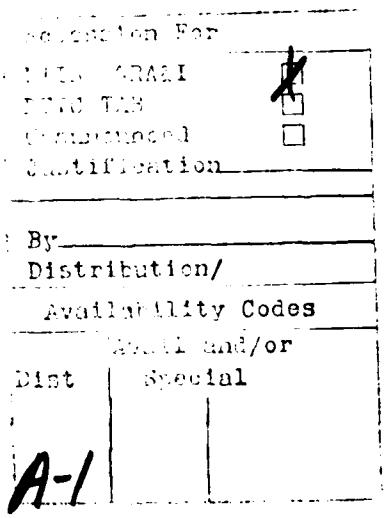
The meaning of a lexical item can be indicated by a pointer to a concept in a semantic network or by a pattern which matches a piece of conceptual structure. Some systems additionally have a concept of lexical choice, i.e., routines which explicitly choose between alternative lexical realizations of a particular meaning.

ISI Research Report
ISI/RR-86-168
October 1986

University of Southern California

Susanna Cumming

The Lexicon in Text Generation



INFORMATION SCIENCES INSTITUTE

213/822-1511

4676 Admiralty Way/Marina del Rey/California 90292-6695

This research is supported by the Defense Advanced Research Projects Agency under Contract No MDA903 81 C 0335 Views and conclusions contained in this report are the author's and should not be interpreted as representing the official opinion or policy of DARPA, the U S Government, or any person or agency connected with them

Table of Contents

1. Introduction[1]

This report reviews the state of the text generation lexicon. I have two primary goals: 1) to give the reader an idea of what is currently being done, by setting out some of the alternatives that designers of generation lexicons have faced, the choices they have made, and the implications of these choices for the types of lexical phenomena they have been able to represent; 2) to suggest what a generation lexicon *could* do, i.e., what range of lexical phenomena is relevant to the generation task. These issues will be addressed more or less in parallel throughout this report, with more attention to the first goal in the first two sections, and to the second in the last three sections.

There are many aspects of lexical representation which I have chosen not to cover in this report. I haven't given much space to a description of morphological information, because most of the systems I have investigated generate English, which isn't very interesting from a morphological point of view. Also, since I haven't looked at any systems which generate speech, there is no discussion here of how to represent phonological or phonetic information.

1.1. What is lexical knowledge?

A brief examination of a few text generation systems reveals what seem to be staggering differences in the content of the component labelled "lexicon" or "dictionary". Treatments range from dictionaries which contain only information about the endings of nouns and verbs, to systems which store entire sentences as single units in the lexicon; from systems which insert lexical material as a last stage in the derivation process, to systems with lexicons that do the major part of structure-building work. However, this apparent diversity is to a large degree illusory: systems represent the same basic kind of information in different ways and in different components. For instance, information about restrictions on the modifiers a word can take can be treated as part of syntax, as part of semantics, or as a purely idiosyncratic component of a

---

[1]This paper has benefitted immeasurably from interaction with a number of my colleagues, most notably Bill Dolan, Cece Ford, Bob Ingria, Johanna Moore, Lynn Poulton, and Sandy Thompson. Special thanks are due to Christian Matthiessen, who has done his best to educate me in a number of areas of general linguistics and text generation where I am deficient. Any misconceptions or inadequacies that remain are my own.

lexical entry. This diversity has its origin in the diversity of practical goals and theoretical underpinnings of the text generation systems I studied.

The diversity of approaches to lexical representation in linguistic theory is not just an artifact of notational differences; it in turn stems at least partly from the fact that the appropriate characterization of a "word" is different in different subsystems of language. In other words, "word" must be differently defined for the purposes of phonological, orthographic, morphological, syntactic, and semantic regularities, although there is a partial overlap (which accounts for the fact that we can frequently get away with using the same term for all these different units). For most of the systems discussed in this report, the only crucial mismatches are those between the syntactic word and the semantic word (though the orthographic word and the morphological word do occasionally have to be dealt with as well).

Because of this complexity, for the purposes of this report I will avoid answering in any absolute way the question posed in the title of this section. Instead, I will characterize as "lexical knowledge" that knowledge which at least one of the systems which I review contains in a component called a "lexicon" or "dictionary". My discussion will principally concern the connections between the structure of particular systems and the decisions made in those systems about whether and how to represent particular pieces of lexical information.

1.2. Understanding vs. generation: different priorities

Before I begin, I would like to address the issue of the extent to which the directionality of linguistic processing -- that is, whether it is a matter of understanding or generation -- influences the content of the lexicon. According to one ideal, in which the language processing system models all of the linguistic knowledge of a human speaker, the relevant information should be the same; and some systems which are bidirectional[2] use the same lexicon for both understanding and generation. However, in practice the two types of lexicon tend to be rather different in the information they

[2]For example, JANUS, the VIE-LANG system, and PHRED. (References for these and the other systems mentioned in this paper are all given in Section 1.3, below.)

encode; even in the bidirectional systems, some of the lexical information is used in only one direction. This is due to differences in the type of demands that apply to most actual understanding and generation projects.[3] A text understanding system must be able to accept whatever input it receives from the user; this requirement dictates a grammar which is comprehensive at least with respect to a given domain, and a dictionary which is both lexically comprehensive (contains a large number of words) and syntactically comprehensive (supports all the syntactic distinctions that the grammar can make). However, it can assume a fluent and cooperative interlocuter; it doesn't have to weed out input which is textually non-cohesive, unidiomatic, uncooperative, or otherwise "awkward" (with the exception perhaps of gross syntactic ungrammaticality). A generator, on the other hand, doesn't need a full range of syntactic capabilities (one way of saying whatever it needs to say may be enough); nor does it need a very large lexicon (one word for each thing it needs to say, and fewer syntactic distinctions corresponding to a smaller syntactic component). But it has to know more about the syntax and lexicon it does have: it must have a basis for choosing between syntactic alternatives and lexical items so as to be not only conceptually appropriate and grammatical, but also cooperative, idiomatic, non-redundant, and otherwise fluent.[4] Thus, we can say that the generation task sets different priorities for the lexicon: roughly speaking, a generation lexicon has to put depth before breadth, while the reverse is true for understanding.

In this report I will naturally concentrate on those aspects of lexical specification which are most particular to the generation task.

---

[3]This remark, as most of the observations in this paper, applies only to natural language systems which are intended to take one side in a communicative exchange with a user. It does not necessarily apply to a system such as ILIAD, which produces sentences for the purpose of language drill, or to a system which generates random sentences in order to test grammar rules.

[4]An analogy can be made to the experience of a human learning a second language: typically the range of the language which the learner can produce appropriately is much smaller than the range the learner can comprehend.

1.3. Systems surveyed

In order to make more concrete the comparison between systems presented in this report, I will first give a very brief sketch of each of the systems I have been able to investigate, with particular attention to the structure and function of the lexicons within the systems. More detailed discussion of the interesting features of various of these lexicons will be given in the body of the report. Citations for the sources from which I have drawn my information are all given in this section; hereafter I will refer to systems by name without repeating the citations. (For the convenience of the reader who may not be familiar with all these systems, I will upper-case system names throughout the text of the report even when this is not the conventional spelling of the system name, so as to distinguish them from the names of the researchers who developed them. In this section, systems are listed alphabetically for easy reference. In some cases, I have assigned a name to unnamed systems.)

I should add that, in most cases, I have not had an opportunity to examine the actual listings for the lexicons I discuss.[5] My statements as to the contents of these listings are inferred from the published descriptions of the systems; frequently only incomplete or suggestive information is provided about the lexicon.[6] Therefore, my comments should be taken as reflecting potential capabilities of particular lexicon formalisms, which may not be fully exploited in the working versions of each system. As the interesting issues have to do with what is possible rather than with what has been done, I don't see this as a liability.

ANA: [Kukich 83a, Kukich 83b]. Generates English text from numerical data about the stock market. The lexicon contains entries for whole subjects and predicates. Each entry contains morphological information, semantic information matching certain patterns in the data, and stylistic information (which aids in lexical selection) as well as lexical material. The predicate entries contain subject slots, with

---

[5]The exceptions are the lexicons of the JANUS system (which I have worked on), TLXT, and ILIAD.

[6]In many systems, especially those with a case-frame orientation, the information available applies only to verb entries; I have much less information about the representation of nouns and even less about other categories.

semantic restrictions on the fillers of these slots.[7] Thus there are predicate entries like "display a hesitant mood early in the day" and "display a hesitant mood late in the day", and subject entries like "the indexes" and "stock indexes".

ILIAD: [Bates & Wilson 81, Bates, Beinashowitz, Brown, Dougherty, Ingria, Shaked, Simpson & Wilson 81, Bates, Beinashowitz, Ingria & Wilson 81, Bates & Ingria 81]. Generates English sentences designed to test language ability in deaf children. The lexicon contains semantic information relating the entry to a conceptual hierarchy, case-frame information with semantic restrictions on the fillers of the slots, and morphological information.

JANUS: [Mann & Matthiessen 83, Matthiessen 84, Cumming & Albano 86, Cumming 86]. A natural language interface which includes the Nigel systemic generation grammar developed at USC/Information Sciences Institute, and the RUS parser developed at Bolt Beranek and Newman, Inc. The parser and the generation grammar share various data structures, including the lexicon. The JANUS lexicon (ML, or Master Lexicon) contains lexical entries which are single words or continuous multi-word phrases. Each entry has a feature specification (which contains morphological as well as syntactic features); a semantic specification, which is the name of one or more concepts in the knowledge base; and possibly some properties which provide cross-indexing with other lexical entries, values for case, and number of pronouns, etc. The features include all the feature information required by the Nigel and RUS grammars; thus some features are used by only one of the grammars. In this discussion my remarks about JANUS feature specification will be aimed primarily at the subset of features used by Nigel.

The features of the Master Lexicon are arranged hierarchically in a tree; they can thus be thought of as defining wordclasses. The wordclass organization contains information about which features are compatible with which other features, and what can constitute a complete feature specification. A word can belong to any number of wordclasses. Thus in some respects the feature hierarchy of the JANUS system is similar to the feature systems represented by the "word ranks" of some other systemic generation grammars (e.g., PROTEUS and SLANG).

---

[7] While there are also slots within predicate entries, these are only for quantitative elements which are inserted from the statistical summary.

KAMP: [Appelt 83, Appelt 85a, Appelt 85b]. Combines a planner with a "teleological grammar" (Telegram) written in Kay's unification framework [Kay 79]. The lexical entries map semantic material to lexical material annotated by syntactic features. Unlike some other grammars written in this framework (e.g., McKeown's grammar), lexical entries apparently do not contain internal structure.

MUMBLE: [McDonald 80, McDonald 83, McDonald 85]. This system produces English text from a variety of input meaning representations. It contains two main knowledge structures, the "dictionary" and the "grammar". The dictionary builds structures by matching an element of the semantic representation to a structure containing lexical material and labelled slots. More than one realization of the semantic representation may be specified, so dictionary entries contain "decision-rules" which choose between alternatives on the basis of context; the various possible outcomes are called "choices". The grammar performs realizations on the structures that emerge from the dictionary.[8]

PHRED: [Jacobs 85, Jacobs 83]. The generation half of a natural language dialogue system; the other part is an analyzer, called PHRAN. The system's principal knowledge structure is the "pattern-concept pair", where the pattern (a phrasal unit which specifies structures, features, and lexical material) is linked to the "concept", a semantic representation; this may be thought of as the lexicon. The same knowledge is used in understanding and generation.

PROTEUS: [Davey 78]. A systemic grammar which generates descriptions of tic-tac-toe games. It treats the lexicon as a "word rank", as proposed in [Halliday 61]; according to this view of lexis,[9] lexical choices are represented exactly as grammatical choices are, as a system network in which each choice has its own "rank". In Davey's system, verbs are treated a little differently: the lexical item corresponding to the verb is chosen within the verbal group rather than in the word rank.

---

[8]The structure of the MUMBLE dictionary seems to have changed somewhat in the version described in [McDonald 85], with the introduction of domain-independent "realization-classes" which contain some of the more general decision-rule choice correspondences and which can be referred to in dictionary entries.

[9]Systemic linguists prefer the term "lexis" to "the lexicon", since the latter term evokes images of a single repository of lexical information which is organized around words rather than choices; I'll discuss this distinction further in Section 2.3.

For convenience, there is also a "lexicon proper" which contains morphological information about lexical items that inflect.[10]

SLANG: [Patten 86]. Another systemic grammar, which generates from a systemic semantic stratum. Like PROTEUS, it represents lexical distinctions in a word rank of the grammar. However, SLANG handles inflected forms as separate words in the grammar, rather than storing inflectional information in a separate component and doing morphology via a routine.

SMRAD: [Kittredge & Mel'chuk 83]. A proposed system which would incorporate the ideas on dictionary content represented in [Mel'chuk 81, Mel'chuk et al. 83, Mel'chuk & Zholkovsky 84]. In addition to semantic, syntactic (including case frames), phonological, and morphological information, a lexical entry contains *lexical functions* which relate the word being defined to other words that conventionally cooccur with it or have certain other types of semantic relationship with it.

TEXT: [McKeown 83, Derr & McKeown 84, McKeown 85]. Generates English text in response to user questions about the structure of a database. The system consists of several components, of which the most important are the strategic component (which creates strings of propositions by selecting a schema and filling it with propositions from the knowledge base with guidance from focus constraints), the dictionary, and the "tactical component", which contains a unification-style grammar and some realization routines. The "dictionary" is intermediate between the strategic component and the unification-style grammar; it matches semantic predicates to verb entries containing lexical material and argument structures, and fills in the arguments from entries corresponding to the arguments of the semantic representation. The grammar performs transformations and syntactic realization on the output of the dictionary. There is also a "lexicon", which contains morphological information used in realization.

VIE-LANG: [Buchberger, Steinacker, Trappl, Trost & Leinfellner 82, Steinacker & Buchberger 83, Steinacker & Trost 83]. A bi-directional German dialogue system; the lexicons (of which one contains morphological information, and the other contains syntactic/semantic information) are shared between the parser and the generator. The syntactic

---

[10]When writing about English, I use the term "inflection" to refer to the addition of endings to nouns, verbs, and adjectives to indicate number, tense, person, and degree.

lexicon contains pairs (similar to the "pattern-concept pairs" of PHRED) which match semantic representations to syntactic patterns including lexical material and case structures.

GAT [Gross 84, Danlos 84, Danlos 85]. (As far as I can tell, this system is unnamed; I've given it the acronym GAT from the name of [Danlos 85].) Generates reports of terrorist attacks in English and French, from summaries of the attacks. It uses the lexicon/grammar developed by M. Gross and others at the LADL project in Paris: the lexicon can be thought of as a list of all the "simple sentences" which exist in the language, with labelled slots for the noun phrase arguments. The "simple sentences" have features specifying the transformations they can undergo, characteristics of the arguments that can fill the slots, etc. These "simple sentences" are such things as "ACTOR explode EXPLOSIVE in VICTIM'S:LOCATION", or "ACTOR open fire on VICTIM'S:VEHICLE".[11]

2. Phrasal Lexicons and Word-Based Lexicons

The lexicons used in text generation systems can be roughly grouped into two classes, according to what is represented in a typical lexical entry (unit of the lexicon). One class contains lexicons whose entries are typically single words, like the lexicons of traditional linguistic theory; the other class contains lexicons whose entries typically represent larger constituents, phrases or even sentences, with some lexical material (by which I mean orthographically realized words which will appear in the output string), and usually also some slots or variables which can be instantiated with further lexical material or lexical entries. The distinction between these two types isn't always clear-cut. Some systems, as mentioned above, have both types, in which typically the phrasal lexicon represents syntactic and semantic information, and the word-based lexicon

[11] My translation of Danlos' examples. The upper-case words are the slots, which are filled in from the event summaries.

represents morphological information;[12] others can easily provide either type of representation, and the alternative chosen in any given case depends upon the researcher.

2.1. Phrasal lexicons

Perhaps the most important factors distinguishing generation lexicons are the size of the lexical item, the amount of structure it contains, and the role of lexical selection in the system. In text generation, as opposed to understanding, there seems to be a tendency towards a large size, a complex structure, and a powerful role for the lexical item. In this section, I will discuss the reasons for each of these tendencies and their implications for text generation; in Section 3, I will describe how more traditional word-based lexicons handle the same range of phenomena.

2.1.1. Size

While traditional dictionaries are primarily organized around small linguistic units -- words or even morphemes -- many computational lexicons have entire syntactic constituents stored as their basic unit, all the way up to multi-clausal units. (These lexicons can conveniently be described as "phrasal", although the kind of unit which counts as a "phrase" varies widely. An argument for this treatment can be found in [Becker 75].) This practice has several advantages in text generation:

1. All kinds of subcategorization and selectional restrictions which need to be stated as properties of particular lexical items can easily be handled without any special mechanism: the allowed patterns are listed in the lexicon, and the disallowed patterns aren't. Any combination of complement types may be represented without the necessity of deciding beforehand on a particular inventory of possibilities.

---

[12] This style of representation is much more efficient where there is a lot of morphological information to be expressed, since in most systems different senses of the same (orthographic or phonological) word will receive different lexical entries, but the inflection will be the same. For example, *be* as a passive auxiliary (as in *the bug **was** eaten by the bat*) and *be* as a copula (as in *the bug **was** a spider*) are very different syntactically and semantically, but they share the same inflected forms (i.e. *am, are, is, was, were, been, being*), as do all the other uses of the verb spelling *be*. If a morphological and a syntactic/semantic lexicon are distinguished, the information about the forms of *be* can be represented only once. In English, the amount of inflectional information that needs to be specified is so small that this may not be an important consideration (*be* is an extreme example), but in other Indo-European languages it becomes much more important.

2. Similarly, all kinds of idioms and collocational restrictions can potentially be handled by specifying the exact wording of the lexical phrase.

3. An indefinitely large syntactic range may be "simulated" by treating as idioms those syntactic constructions which can't be generated by the grammar, thus adding to the syntactic variety of the output text. This principle may be extended to the point where the lexicon "takes over" most of the grammar, i.e., all or almost all grammatical patterns are represented only in the specification for the lexical items to which they apply.

The disadvantages of this method are merely the flip side of the advantages. Generally speaking, the more phenomena represented as idiosyncratic properties of lexical items, the fewer phenomena are treated in a general way (although some systems have the flexibility to represent the same phenomena as either idiosyncratic or general). This has two related consequences: 1) lexicons must be much larger; 2) making additions to the lexicon is a much more lengthy and difficult process, as properties of lexical items which may in fact be predictable (on the basis of other lexical properties or semantic properties of the item) must be specified anyway.

2.1.2. Structure

Phrasal lexicons differ in the amount of internal structure they can encode within their phrases. Thus, there is a difference between encoding an idiom like *go mad* as a verb or predicate with no internal structure indicated and knowing that *go* is a verb and *mad* is a resultative adjective phrase. If internal structure is indicated, it is possible to store each of these variants as a single lexical item (which may be desirable, since the phenomenon is not generally productive), and yet still allow some syntactic variation, e.g., adding intervening adverbials (*go quietly mad*), inflecting the verb (*I go mad, he goes mad*), or relating the idiom to other syntactically similar expressions (*go crazy, run dry*). Information about the internal structure of phrases is also necessary for stylistic control, e.g., to allow control of the amount of variation in lexical choice and syntactic structure.[13] The lexicons of TEXT, PHRED, VIE-LANG, and MUMBLE all allow any amount of internal structure to be specified in a lexical item, in contrast to GAT and ANA; while these two systems contain slots for other elements (various

[13]Kukich discusses this point in [Kukich 83b], p. 124.

arguments in GAT, subjects only in ANA), they cannot indicate any further structural complexity.

2.1.3. Depth of lexical selection

Another important parameter which distinguishes generation lexicons is the amount of influence lexical choice has over other kinds of choices (for example, syntactic, rhetorical, or stylistic choices) made in the system. Lexical choice can often restrict clause syntax: for example, some verbs with direct objects can't be passivized (*The candy bar cost a quarter*); verbs (and, to a lesser degree, adjectives and nouns) restrict the syntax of their complement clauses in various ways (*I insist that he come* vs. **I insist that he comes*, but *I hope that he comes* vs. **I hope that he come*); some pronouns can be modified by relative clauses while others can't (*Anyone who wants to can come* but **We who want to can come*[14]). Naturally, the degree of constraint that availability of lexical items can impose on grammatical choice is directly related to the stage in the generation process (or "depth", in terms of the metaphor current in transformational grammar) at which lexical choice is made. If lexical choice is made late in the generation process, it can have little input into other decision-making, unless some kind of backtracking is allowed.

In many systems, the lexicon acts as the intermediary between semantic and syntactic representations, and the step of "lexical insertion" is actually the step at which syntactic structure is built. (This is the case for MUMBLE, TEXT, PHRED, VIE-LANG, and GAT.) This generally works by matching the predicate of the semantic representation with the lexical entry for a verb, and then filling in the argument slots of the verb with arguments from the semantic representation. (It may also be more complicated than this: in both TEXT and MUMBLE, for instance, the way this matching is done may involve information from contextual information such as focus history or preceding reference; and in ANA, stylistic factors such as length are considered.) In these systems, the structure built by the lexicon then undergoes further syntactic realization (e.g., transformations, morphological adjustments). Since the

---

[14]The latter example may be grammatical with a nonrestrictive reading, but it is not possible with a restrictive reading.

lexical item has already been chosen when these realizations are performed, properties of the lexical item have the opportunity to constrain the way these realizations occur. For example, in TEXT, routines in the dictionary itself control the choice of syntactic construction (active, passive, or existential) as well as the basic sentence structure. This avoids problems such as a text plan calling for passive syntax when the verb in question can't be passivized. In KAMP, syntactic processing (including lexical insertion) is alternated with planning in such a way that plans can be modified in response to the set of choices made available by a particular lexical item. In GAT, all the decisions are made simultaneously by the selection of a particular schema which includes lexical, (clause-level) syntactic, and clause-combining specifications.

Of course, if a grammar is sufficiently rich to treat as regular (i.e., as predictable from aspects of the specification of the sentence) a large range of syntactic phenomena, a correspondingly small range needs to be treated as idiosyncratic to a lexical item (i.e., as dependent on a particular lexical choice). This is another form of the tradeoff between grammar and lexicon: the more complete a grammar is, the less dependent it is on early lexical specification to do its job properly. Thus, in Nigel, most of the syntactic properties of a lexical item are taken to be predictable from its semantic properties, following Halliday's analysis; so, although a particular lexical item isn't chosen until after syntactic planning has occurred, the syntactic plan is made with reference to the same semantic categories that constrain lexical choice.[15] For example, non-subjunctive "that" clauses, since they refer to reports about the world, are restricted to verbs of saying and thinking.

2.2. Word-based lexicons

Until recently, many of the models of language to come out of linguistics have assumed a word-based lexicon in which syntactic information is specified in the form of features. In this type of lexicon, word choice has been constrained on the basis both of meaning and of the fit between the syntactic features of the word and the syntactic

---

[15]Of course, this statement is relative to a particular view of the characterization of both syntax and semantics; for more discussion of this point, see Section 4 below.

environment into which it is supposed to fit. Rather than having the powerful role it has in the systems discussed above, the lexicon has been viewed primarily as an appendage to the syntax, where information which can't be predicted by general rules is stored. The units represented have been small (usually morphemes), and the amount of internal structure which can be represented within an item has been minimal. Systems surveyed here which have this traditional type of lexicon are ILIAD, KAMP,[16] and Mel'chuk's system, SMRAD.[17]

In some ways, the difference in practice between a low-level word-based lexicon with features and a highly structured phrasal lexicon is smaller than it appears. For example, a case-frame representation can be mapped onto a feature representation in which the feature corresponds to a particular case pattern -- e.g., the feature "transitive" can be mapped onto a case frame containing a direct object slot. The major difference is that the case frame representation allows more freedom than is available with a small set of features (as mentioned above); on the other hand, since features can be thought of as corresponding to classes of lexical items, a single lexical feature may efficiently encode a range of possible case frames that tend to cooccur with a particular type of word. In the lexical feature specifications referred to by Nigel, all of the subcategorizational possibilities of a particular sense of a verb are taken to be predictable from a single feature representing its wordclass membership.[18] Thus, verbs such as "see" and "hear" have the feature "perception"; the grammar knows that these verbs can be generated either with a direct object, with a complement clause in which the verb is in its stem form without "to" (e.g., "I saw you arrive", "I heard her come

---

[16]Although the Unification formalism used in KAMP allows for lexical entries containing further structure, just as in Lexical Functional Grammar representations, as far as I know Appelt doesn't exploit this possibility in his system.

[17]Although Mel'chuk's dictionaries contain an unusual degree of cross-referencing between entries, they are still primarily organized around entries for single words.

[18]These features are related to the semantic type of the verb as represented in the position of the corresponding concept in the semantic network; however, the relationship is not direct. As we will see in Section 3.2 below, "deep case" phenomena and selectional restrictions are also handled in the JANUS system; however, they are treated purely as part of knowledge about word meanings, and therefore represented in the semantic net rather than the lexicon.

in"), or with a complement clause in which the verb is in its present participle form ("I saw you arriving", "I heard her coming in"). This particular configuration of possible complements is restricted to verbs that refer to sense perception, and thus it is redundant to list each of these possibilities separately for all the perception verbs.

Of course, to take advantage of this type of generalization one must have an account of the wordclasses of a language, as reflected in both semantic class and argument structure; and indeed, it's clear that a reasonably complete grammar must make reference to a very large set of such wordclasses. This is another case of a tradeoff between having a relatively complex rule system that treats few things as "irregular" or unpredictable, and having a relatively simple rule system that treats many things as irregular. In the computational context, the first option implies a large development effort in the area of grammar, while the second implies a large effort in the area of lexicon. The goals of the system determine which option is preferable.

2.3. Systemic grammars

The systemic approach to lexical classification, exemplified in SLANG and PROTEUS, doesn't fall easily into either of the categories described above, although in practice these two systems, like Nigel, have the closest affinity with word-based systems, since neither supports phrasal lexical items.

The "word rank" of a systemic grammar represents alternatives among wordclasses in the same way the grammar represents grammatical alternatives; the result is a highly structured feature system. Within the word rank, successive choices lead to actual words in the case of closed-class items or "function words" such as prepositions, verbal auxiliaries, and connectives; these can be thought of as words with unique feature specifications. As mentioned above, the wordclass hierarchy of JANUS is similar in some ways to a word rank; however, it is more limited in the kinds of relationships it can represent between features.

In systemic theory, choices between open-class items fall into the area called "lexis", often envisioned as an entirely separate level of grammar [Halliday, McIntosh, &

Strevens 64, Berry 77, Halliday 76]. It has been proposed that lexis could ultimately be entirely incorporated into the grammar -- that is, that finer and finer (or, as systemicists say, "more and more delicate") decisions could ultimately distinguish every word from every other word -- but this "dream" (as Halliday has called it [Halliday 61]) has never been completely realized.

3. Approaches to Cooccurrence Phenomena

Now that we have surveyed the various kinds of lexicon and the way they interact with the systems of which they form a part, we will examine the range of phenomena that they express, and consider the implications of these phenomena for optimal lexicon design. Most of the syntactic information (and some of the semantic information) that needs to be specified about lexical items can be subsumed under the term "cooccurrence information", i.e., information about the other linguistic elements (lexical items or syntactic types) that can "go with" a particular item. I will discuss here four distinct types of cooccurrence phenomena: subcategorization, selectional restrictions, collocation, and idioms.[19] By "subcategorization" I mean specification of the syntactic or semantic frame(s) in which an item can occur, such as the fact that *think* can take a clausal complement with *that* but not a complement with *to*. By "selectional restrictions" I mean semantic restrictions on the fillers of subcategorization frames, such as the restriction on the subject of the verb *elapse* that it refer to a period of time. By "collocation" I mean lexical restrictions (restrictions which are not predictable from the syntactic or semantic properties of the items) on the modifiers of an item; for example, *answer the door* is acceptable, but **answer the window* is not. By "idiom" I mean a fixed phrase whose meaning is noncompositional, i.e., not predictable from the meanings of its parts, for example, *a one-track mind*; an idiom may be "ungrammatical" (i.e., not generatable by independently motivated rules) if interpreted compositionally, for example, *all of a sudden*.

---

[19] My use of the terms "subcategorization" and "selectional restriction" is largely derived from their use in classical transformational theory. "Collocation" in this sense can be traced back to [Firth 57]; my sense is related most specifically to Firth's "general or usual collocations". "Idiom" as used here is more restricted than the sense it is given in, for example, *Longman Dictionary of English Idioms* [Longman 79] (which includes collocations, standard metaphors, proverbs, etc., as well); it is closer to what are characterized as "traditional idioms" in the introduction to Longman's.

A consideration of these definitions will at once suggest that the extension of these classes of phenomena depends largely on the particular model to which they are applied. Whether something needs to be treated as compositional or not will depend on the rules that are available to generate it; there are large numbers of constructions which apply to very limited classes of words. For example, there is a set of expressions *hundreds and hundreds, thousands and thousands*, etc; this construction is limited to number words that act like common nouns in that they can be plural and take articles (so we get *a dozen, several dozens, dozens and dozens* but not **a twelve, *several twelves, *twelves and twelves*), and also to other kinds of quantity expressions, e.g., *barrels and barrels.* While this could be treated as a regular grammatical construction, it is sufficiently limited in generality that few computational grammars will include it in their syntactic scope; it may be more cost-effective to treat this kind of phenomenon as idiomatic. Similarly, what could be stated as a selectional restrifction (if the right semantic classes were added to the model) may otherwise have to be stated as a set of collocations or idioms. And the line between selection and subcategorization is blurred when syntactic properties are taken to be predictable from semantic classes.

3.1. Subcategorization

The handling of subcategorization in several models has been touched on above, in Section 2.1.1. To reiterate, most phrasal or case-frame lexicons indicate subcategorization by using slots in a lexical entry. The following lexical entry from PHRED ([Jacobs 85], p. 221) for the verb *remove* is fairly representative:

<agent> <root = remove> <physob>
<<word = from> <container>>

This entry contains the information that the verb "remove" takes a subject (which is an agent), a direct object, and prepositional phrase with *from*. (It also places certain semantic restrictions on the fillers of these slots.)

Word-based lexicons, on the other hand, generally deal with subcategorization by providing lists of features. The entry from the JANUS lexicon for the same verb contains the following syntactic (and morphological) information:

```
(make-lexical-item
  :name 'REMOVE
  :spelling "remove"
  :features '(VERB INFLECTABLE UNITARYSPELLING S-D LEXICAL
CASEPREPOSITIONS OBJECTPERMITTED PASSIVE DOVERB
DISPOSAL EFFECTIVE) )
```

3.2. Selectional restrictions

Some lexicons can handle selectional restriction by attaching semantic restrictions to lexical entry slots. The labels *agent*, *physob*, and *container* in the PHRED example above can be thought of as selectional restrictions. ILIAD lexical entries contain similar restrictions; for example, the entry for (the verb) "grease" is as follows:

```
(GREASE SYNCASES
 ((SUBJ (HEADCONCEPT T) (MUST-BE (OR (ADULT CHILD))))
  (OBJ (HEADCONCEPT T) (MUST-BE VEHICLE))))
```

This says that the subject of "grease" must be a word that refers to an adult or a child, while the object must refer to a vehicle. ANA's predicates contain feature restrictions on their subjects (e.g., the entry for *display a hesitant mood early in the day* has the features **^subjtype NAME ^subjclass MKT**, indicating that the subject must be a name for the stock market), and the slots in the "simple sentence" lexical items of the LADL grammar may have semantic feature restrictions such as +HUMAN associated with them.

In other lexicons, including those of JANUS and TEXT, selectional restrictions aren't directly represented in the lexicon at all; rather, these restrictions are in fact captured in another part of the system -- the semantic network. This option is available to systems that are based on semantic networks composed of hierarchically arranged concepts, related to one another by "case roles" (which specify the semantic roles a concept has and the other concepts that represent possible fillers of each role). In systems that use a semantic net as the source of the representations which go to the grammar, selectional restrictions are already enforced in the representation that goes to the grammar for expression. This is equivalent to saying that selection, unlike subcategorization, derives

from knowledge about the meanings of words rather than lexical knowledge specific to the linguistic expressions of those meanings.[20]

3.3. Collocation

The phenomenon I have called collocation is of particular interest in the context of a report on the lexicon in text generation, because this particular type of idiom is something which a generator needs to know about, while a parser may not. For example, consider the expression *wreak havoc*. This can be parsed compositionally as a verb and its object without any special knowledge; but a generator must know about the special connection between these words, since neither word is found very often in any other context; we need to avoid generating *wreak a mess*, *make havoc*. (Many more examples of this kind of expression can be found in [Makkai 72, Chafe 68, Fillmore 79, Fillmore, Kay & O'Conner 84].) Because of this, this set of phenomena has been labelled "idioms of encoding",[21] i.e., expressions which are compositional, and may seem semantically transparent to a hearer but require specialized knowledge on the part of a speaker to produce correctly; non-compositional cooccurrence phenomena like *kick the bucket*, the ones which I call "idioms" here, correspond to Fillmore's "idioms of decoding"; both a parser and a generator must have knowledge of these.

Collocation phenomena aren't explicitly handled as such by any of the systems discussed so far.[22] They can, of course, be handled after a fashion, either by treating them as cases of selection (as the JANUS system does) or as cases of idioms (as in the PHRED system). If they are handled as selection, the distinction between idiosyncratic lexical properties and general semantic properties is lost; and if they are handled as idioms, the regular syntactic behaviour and semantic compositionality of these phrases

---

[20]Systems differ, however, in how close the mappings are between concepts and words, semantic role specifications and syntactic case frames; in some systems it would be hard to argue that the properties of the "concepts" of the semantic net aren't simply properties of the words used to express those concepts in a particular language, or that the "semantic roles" on those concepts aren't really labels for syntactic arguments. For more discussion of this issue, see Section 4.

[21]I believe the term comes from [Makkai 72].

[22]While Jacobs discusses these phenomena in [Jacobs 85], he doesn't actually distinguish them from idioms (of decoding) in his system.

isn't expressed. Thus, neither of these solutions is perfectly satisfactory, although one or the other may be adequate for a small domain in which full generality isn't crucial.

The only system I know of which addresses this kind of phenomenon in a thorough and explicit way is that described in [Kittredge & Mel'chuk 83]. They have proposed a device called the "lexical function", which he uses extensively to relate dictionary entries in his "explanatory and combinatorial" dictionaries of Russian and French. There are a large number of these lexical functions (62 "standard" ones, and an arbitrary number of "non-standard" ones), but they can be roughly divided into two groups: those that deal with paradigmatic relationships between words (meaning relationships such as hyponymy, synonymy, antonymy, etc., plus words with related meanings but permuted argument structures; for more discussion of some of these phenomena, see Section 4 below), and those that deal with syntagmatic relations -- standard words for the various arguments and modifiers of a term. It is this latter group of lexical functions that can be taken as expressing collocational phenomena. For example, there is a function *Magn* which relates a word with a modifier which has the meaning "to a great degree"; the words "shave", "easy", "scoundrel" have as Magns "close", "as pie", and "unmitigated", respectively. Presumably these lexical functions will be exploited in the SMRAD text generation system proposed in [Kittredge & Mel'chuk 83].[23]

3.4. Idioms

Idioms have been discussed in some detail in Section 2.1.1 above and in the preceding paragraphs of this section. To reiterate, most phrasal lexicons can generally handle idioms without any special provisions, either by treating all pieces of the idiom as part of the same word, as in Kukich's system, or (in case-frame lexicons) by having some of the slots filled in with lexical material. For example, in PHRED, *tell (someone) to get lost* is

---

[23]Hudson distinguishes idioms and collocations more or less the same way I do here, in his "Word Grammar" theory [Hudson 84]; his theory is actually quite similar to Mel'chuk's dependency grammar. However, as far as I know Hudson has no proposal for a text generator, so a discussion of his account would be out of place here.

<person> <root = tell> <person>
<word = to> <word = get> <word = lost>

(Note that there is relatively little internal structure to this idiom; in particular, "to get lost" is not a clause, or indeed a constitutent.)

In the word-based systems I've surveyed, idioms can only be handled as single words, with no intervening material (thus *kick the bucket* can be handled -- as an intransitive verb -- but *knock (someone's) block off* can't be, and *kicked/kicks the bucket* may or may not be). JANUS can't handle internal inflection, so idioms which are verb phrases aren't possible at all; however, anything that doesn't have to inflect internally is allowed, such as many noun phrase idioms (such as *red herring*, which like other nouns can pluralize by adding an ending to the entire lexical item, as *red herrings*), and such things as complex prepositions (such as *face to face with*, *on account of*) can also be handled.

4. Lexical Semantics and Lexical Choice

If the phenomena treated in the previous section are characterized as phenomena of syntagmatic organization -- i.e., facts about what a lexical item can occur next to -- then the facts discussed in this section can be thought of as facts about paradigmatic organization -- i.e., facts about what a lexical item can occur instead of, or facts about lexical choice and meaning relations between words of the same class. The topic of lexical semantics will be treated only rather briefly in this report (relative, at least, to the amount that has been said about it in the theoretical literature), since not all systems have an identifiable component of lexical semantics -- separate, that is, from whatever organizing principles underlie the elements of the demands for expression that are interpreted by the generator. Similarly, not all systems have an explicit strategy for lexical choice, obviating the need for decision procedures by relying on a one-to-one mapping between items in the lexicon and elements of the semantic representation.

4.1. Semantic classification

The two basic methods by which systems notate semantic classification of lexical items are by feature systems and taxonomies. (While Mel'chuk's paradigmatic lexical functions might appear to represent a third system, they are based on an underlying taxonomy.) The only lexicon which uses a pure feature system is that of ANA: the phrases of ANA's system are represented as feature clusters (or, more accurately, as clusters of attribute-value pairs). For example, the four entries **display a hesitant mood early in the day, display a hesitant mood late in the day, creep upward early in the session,** and **creep upward late in the session**, are distinguished by the values of the two attributes ^**tim** (time) and ^**deg** (degree).

Explicit taxonomic concept hierarchies represent (at least) relations of inclusion among word meanings. Thus, a taxonomy can represent the fact that a cat is a kind of animal; i.e., that the set of cats is included in the set of animals. Taxonomies can also represent the inheritance of properties from more general to less general concepts; thus, if a cat is an animal and an animal can have young, then a cat can have young. Taxonomies are composed of concepts, each of which may be associated with one or more lexical entries; the lexicon is generally the place where the correspondence between concepts and words is stated. In the above example, we can say that the concept associated with the word "cat" is a **subconcept** of the concept associated with the word "animal", and that the concept associated with the word "animal" is a **superconcept** of the concept associated with the word "cat". In the following discussion, I will use upper case for concept names to avoid confusing them with their associated lexical items.

Systems with taxonomies use taxonomic information in radically different ways. In TEXT, a taxonomy is actually the source of the semantic representations (propositions) from which sentences are generated, since the purpose of the generator is to describe the taxonomy. In JANUS, taxonomic information is used in the reasoning performed by the grammar during the generation of sentences. Thus, if the system is generating the sentence "Jones sent the message", the grammar will look at the taxonomy to see if SEND is the kind of process that typically has an agent. In fact, SEND is a subconcept of the concept DIRECTED ACTION, and since the grammar knows that directed

actions have agents it will construct an agent noun phrase. Thus, the taxonomy employed in the JANUS system contains all the category distinctions relevant to grammatical choice.

In ILIAD, since its function is to provide grammar drills, the demand for expression consists of a syntactic form; the semantic taxonomy is used to ensure that the sentence which is finally generated is semantically coherent, i.e., doesn't violate selectional restrictions. Thus, lexical choice is primarily conditioned by selectional restrictions stated in terms of the taxonomy. For instance, in the example in Section 3.2 above, once "grease" had been chosen as a main verb, the only lexical items which would be considered for the direct object would be those associated with subconcepts of VEHICLE. (Since the actual semantic content of the generated sentence is unimportant in ILIAD, once selectional restrictions have been satisfied, lexical choice is essentially random.)

SMRAD contains a richer specification of paradigmatic relations than any of the systems so far discussed. In addition to hyponymy (the relation between a concept and its superconcept), he has functions for different kinds of synonyms and antonyms, words which have the same basic meaning but with the syntactic roles of the arguments interchanged (e.g., "buy" and "sell"), and many others that aren't so easily classifiable. This richness is vital in a system whose primary goal is paraphrase or translation, since it gives the system access to a great deal of knowledge about expressions that can be considered semantically equivalent, something not available from a simple taxonomy.

4.2. Lexical choice

As described above, some systems do all their lexical choice in what might be called the semantics -- that is, by the time they've decided what to say and before they've looked into the lexicon, they've already committed themselves to a particular wording. Systemic grammars conataining a word rank, conversely, treat lexical choice as part of grammatical choice; "grammar" is often referred to by systemicists as "lexico-grammar" for this reason. (Even in JANUS, this approach to lexical selection obtains for function words, since these are uniquely selected in various ranks of the grammar.)

However, some systems have routines for performing lexical choice built into the lexicon itself.

TEXT has choice routines built into the dictionary, but they are limited to choice of syntactic category: a given element in the demand for expression can have lexical realization in more than one category. For example, SURFACE can be realized as "surface" if it is an adjective or a noun, or as "on the surface" if it is a prepositional phrase. MUMBLE's decision rules combine grammatical choices with stylistic choices. ANA's lexicon provides for choosing in order to enhance stylistic variations of various kinds. Each entry is annotated for its length in syllables, and other things being equal, the grammar chooses so as to alternate two long sentences with one short one; similarly, each subject entry is annotated for "hyponym level", so that on the first mention of a given referent a more specific or more heavily modified phrase is used, and on subsequent mentions more general or briefer phrases are used. For example, *the Dow*, *the industrials average*, and *the Dow Jones average of 30 industrials* have successively lower hyponym levels.

5. Some Goals for the Generation Lexicon

In this section I will summarize the directions which have already been touched on for the generation lexicon, and add a few new goals to the wish list. These represent sets of phenomena which system implementors, regardless of the overall design or underlying linguistic framework of the system, might consider handling somewhere in the system. Some of these goals are met in some of the systems described here; others as far as I know have not been adequately dealt with in any working text generation system, and can thus be considered fruitful areas for future research. Many of them will only be relevant in a really comprehensive text generation system, and can easily be ignored in systems which operate in highly restricted domains.

5.1. Syntactic range

This isn't, of course, strictly a lexicon issue, but one that has repercussions for lexicon design. Most current systems are able to give quite detailed specifications for the subcategorizational properties of verbs, but other syntactic categories also impose subcategorization restrictions on their modifiers. For example, nouns and adjectives[24] can take postmodifying clauses with *that* (*the fact **that** the world is round is well known, it's good **that** you could make it*), as can certain verbs. Similarly, all of the systems I researched know about the inflections of verbs (e.g., *run/runs/ran/run/running*) and nouns (e.g., *book/books* or *goose/geese*), and some know about the inflections of adjectives (e.g., *large/larger/largest*), but none that I know of can generate inflected adverbs, which have the same possibilities as adjectives in English (e.g., *He ran fast/faster/fastest*).[25] For complete coverage, these possibilities must be allowed for.

5.2. The intelligent lexicon

It is a common observation that human languages have many words for things that their speakers commonly talk about (cf. the famous claim, attributed to Whorf, that the Eskimos have twenty words for snow). Less universally accepted is the converse claim that people tend to think and talk about things for which their language has many words. Whether or not this is the case, a text generation system should not plan to say things which it cannot produce with existing lexical resources.[26] In order to assure that this does not happen, the lexical resources of a system should be consulted along with the grammar, semantics, and strategic components in planning what to say, so that if it is not possible to say something using a single word, a periphrastic expression can be planned. As mentioned above, work has been done on this problem in JANUS; KAMP

---

[24]These are the nouns and adjectives that refer to or are predicated on reports of states of affairs; hence the term "factive", which is sometimes applied to them.

[25]Systems also differ as to whether every inflected form must be listed for every inflectable word or phrase, or whether some cases are treated as predictable.

[26]This is not an uncontroversial statement: [McDonald 80] and [Kukich 83b] both argue that the fact that their systems are occasionally "at a loss for words" -- i.e., break down due to the absence of lexical material for something they have committed themselves to express -- is a positive feature, since it accurately models the behaviour of the human language user.

and MUMBLE also both allow for some interaction between planning and linguistic realization such that this kind of negotiation is feasible.

5.3. Cooccurrence phenomena

Ideally, a text generation system should be able to handle all of the phenomena discussed above -- subcategorization, selectional restrictions, collocation, and idioms -- in such a way that the different degrees of productivity and the different restrictions on these phenomena are distinguished. Moreover, the ideal system should have the flexibility either to treat grammatical idioms and grammatical "fixed expressions" productively (i.e., generate them according to general rules), or to store them as units for the sake of efficiency, depending on the requirements of a given domain. Thus, for example, the phrase *We must conclude that...* can be stored as an idiom with a sense equivalent to "therefore", or generated "from first principles" as a clause with a first person plural subject, a modal of necessity, etc. In such a system the tradeoff between productive capability and efficient processing could be avoided, much the way it presumably is in human language use.

5.4. Metaphor

A large range of phenomena which have been treated as idiosyncratic to individual lexical items -- i.e., as idioms or collocations -- could perhaps be treated in a more motivated way in a system which had a notion of standard metaphor. (This proposal is cogently stated in [Jacobs 85]; the sense of metaphor involved here is that presented in, for example, [Lakoff & Johnson 80].) Consider the metaphor "time is money". In a system which had a way of representing this association, a number of collocations involving time ("spend time", "waste time", "lose time", etc.) are not random, but can be predicted from the corresponding collocations involving money. Another set of expressions involving time ("time passed", "time flies", "the days marched by in weary succession", etc.) are derived from another standard metaphor for time, namely, "time is a moving object". While some of Mel'chuk's lexical functions have to do with standard metaphors of this sort, as far as I know his is the only system that treats them systematically as such, although any system based on a taxonomic hierarchy with inheritance can simulate metaphor after a fashion. For example, the popular metaphor

"a computer is a conscious being" is involved when we refer to computers as agents of processes that normally take only conscious agents, e.g., "the computer deleted my files". In the Janus system, the only convenient way to represent this is by classifying the concept COMPUTER under CONSCIOUS BEING in the semantic taxonomy. Ideally, however, it would be preferable not to commit one's taxonomy to the claim that a computer is literally a conscious being, since we also talk about computers as unconscious objects; e.g., we usually say "the computer that just went down", not "the computer who just went down".

5.5. Choice

Ideally, a system should have some way of choosing between lexical items on other than purely grammatical and denotational grounds. Human speakers take a variety of factors into consideration when making lexical decisions. We use different words for the same things, depending on who we're talking to, what we're talking about, where we are, and what role we're playing. A simple example is the observation that in more formal contexts English speakers tend to use Latinate words such as "expunge, remove, infer" instead of Anglo-Saxon phrasal verbs like "wipe out, take off, figure out". In addition to simply responding to social context in the way we choose words, we can use words in a way which evokes or creates a context for our utterances; for instance, we can use borrowings from French in order to sound suave, or surfer slang in order to sound cool. We use more general or more specific terms for the same thing, depending on which of its characteristics we're interested in: if we see a friend careening towards a tree, we're more likely to say "watch out for that tree!" than "watch out for that eucalyptus!" or "watch out for that plant!", because it's relevant that the object is a large and woody plant, but the type of bark and shape of the leaves are irrelevant. And so on. We're a long way from having natural language generators that have the degree of control over any level of linguistic choice, grammatical or lexical, that a serious treatment of these considerations would entail; but we can design our systems so that sucxh distinctions will be able to be accommodated when we have the analyses to support them.

5.6. Conclusion

Lexicons play a wide variety of roles in text generation systems, from the very central one of providing the primary link between form and meaning, to the quite peripheral one of finishing up after the grammar is done. Lexical phenomena such as semantic relationships, syntactic classes, collocation, and idioms have received vastly different amounts of attention in different systems, while other phenomena, such as metaphor and non-denotational meaning, have received virtually none in any system. Examining the capabilities of a wide range of generation lexicons provides an exhilarating sense of the potential for future systems, both from the variety of phenomena dealt with by existing systems, and from the challenges that still remain. I hope that bringing a few of these phenomena to light in this report will succeed in sparking the interest necessary to ensure the lexicon the attention it warrants in text generation research.

References

[Appelt 83] Douglas E. Appelt, "Telegram: a grammar formalism for language planning," in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 595-599, IJCAI, Aug 1983.

[Appelt 85a] Douglas E. Appelt, *Planning English Sentences*, Cambridge University Press, Cambridge, 1985.

[Appelt 85b] Douglas E. Appelt, "Planning English referring expressions," *Artificial Intelligence* 26, 1985, 1-33.

[Bates & Ingria 81] M. Bates, J. Beinashowitz, R. Ingria, & K. Wilson , "Controlled Transformational Sentence Generation," in *Proceedings of the 1981 Meeting of the Association for Computational Linguistics*, ACL, 1981.

[Bates & Wilson 81] Madeleine Bates and Kirk Wilson, *ILIAD: Interactive Language Instruction Assistance for the Deaf*, BBN, 10 Moulton St., Cambridge, MA 02138, Technical Report 4771, Sep 1981.

[Bates, Beinashowitz, Brown, Dougherty, Ingria, Shaked, Simpson & Wilson 81] M. Bates, J. Beinashowitz, D. Brown, D. Dougherty, R. Ingria, V. Shaked, W. Simpson, & K. Wilson, *ILIAD Database Reference*, BBN, 10 Moulton St., Cambridge, MA 02138, Supplement to Tech Report 4771, Sep 1981.

[Bates, Beinashowitz, Ingria & Wilson 81] M. Bates, J. Beinashowitz, R. Ingria, & K. Wilson , "Generative Tutorial Systems," in *Proceedings of the 1981 Meeting of the Association for the Development of Computer-Based Instructional Systems*, 1981.

[Becker 75] Becker, J.D., "The phrasal lexicon," in Schank & Webber (eds.), *Theoretical Issues in Natural Language Processing*, Cambridge, 1975.

[Berry 77] M. Berry, *Introduction to Systemic Linguistics*, Batsford, London, 1977.

[Buchberger, Steinacker, Trappl, Trost & Leinfellner 82] Ernst Buchberger, Ingeborg Steinacker, Robert Trappl, Harald Trost, Elisabeth Leinfellner, "VIE-LANG: A German Language Understanding System," in *Cybernetics and Systems Research*, pp. 869-874, North-Holland, Amsterdam, 1982.

[Chafe 68] Wallace Chafe, "Idiomaticity as an anomaly in the Chomskyan paradigm," *Foundations of Language* 6, (1), 1968.

[Cumming 86] Susanna Cumming, *Design of a Master Lexicon*, USC/Information Sciences Institute, Technical Report ISI/RR-85-163, Feb 1986.

[Cumming & Albano 86] Susanna Cumming and Robert Albano, *A guide to lexical acquisition in the JANUS system*, USC/Information Sciences Institute, Technical Report ISI/RR-85-162, Feb 1986.

[Danlos 84] Laurence Danlos, "Conceptual and linguistic decisions in generation," in *Proceedings of Coling84*, pp. 501-504, COLING, July 1984.

[Danlos 85] Laurence Danlos, *Generation automatique de textes en langues naturelles,* Masson, Paris, 1985.

[Davey 78] Anthony Davey, *Discourse Production,* Edinburgh University Press, Edinburgh, 1978.

[Derr & McKeown 84] Marcia A. Derr and Kathleen R. McKeown, "Using fucus to generate complex and simple sentences," in *Proceedings of Coling84*, pp. 319-326, COLING, July 1984.

[Fillmore 79] Charles Fillmore, "Innocence: a second idealization for linguistics," in *Proceedings of the 5th Annual Meeting of the Berkeley Linguistics Society*, BLS, 1979.

[Fillmore, Kay & O'Conner 84] Charles Fillmore, Paul Kay & M.C. O'Conner, *Regularity and idiomaticity in grammar: the case of let alone*, University of California, Coginitive Science Working Paper, 1984.

[Firth 57] J.R. Firth, *Modes of Meaning*, Oxford University Press, Oxford, , 1957.

[Gross 84] Maurice Gross, "Lexicon-grammar and the syntactic analysis of French," in *Proceedings of Coling84*, pp. 275-282, COLING, Jul 1984.

[Halliday 61] M.A.K. Halliday, "Categories of the Theory of Grammar," *Word* 17, 1961.

[Halliday 76] Halliday, M.A.K., "Lexical Relations," in G.R. Kress (ed.), *Halliday: system and function in language*, Oxford University Press, London, 1976.

[Halliday, McIntosh, & Strevens 64] M.A.K. Halliday, Angus McIntosh, & Peter Strevens, *The Linguistic Sciences and Language Teaching,* Indiana University Press, Bloomington, 1964.

[Hudson 84] Richard Hudson, *Word Grammar,* Blackwell, Oxford, 1984.

[Jacobs 83] Paul S. Jacobs, "Generation in a natural language interface," in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 610-612, IJCAI, Aug 1983.

[Jacobs 85] Paul S. Jacobs, "PHRED: a generator for natural language interfaces," *ACL* 11, (4), 1985, 219-242.

[Kay 79] Martin Kay, "Functional Grammar," in *Proceedings of the 5th Annual Meeting of the Berkeley Linguistics Society*, pp. 142-158, BLS, 1979.

[Kittredge & Mel'chuk 83] Richard Kittredge & Igor Mel'chuk, "Towards a computable model of meaning-text relations within a natural sublanguage," in , pp. 657-659, IJCAI, 1983.

[Kukich 83a] Karen Kukich, "Design of a knowledge-based report generator," in *Proceedings of the 21st Annual Meeting*, ACL, Jun 1983.

[Kukich 83b] Karen Kukich, *Knowledge-based report generation*, Ph.D. thesis, University of Pittsburgh, Interdisciplinary Department of Information Science, Aug 1983.

[Lakoff & Johnson 80] George Lakoff & David Johnson, *Metaphors We Live By,* University of Chicago Press, 1980.

[Longman 79] Longman Group Ltd., *Longman Dictionary of English Idioms,* Longman, Harlow and London, 1979.

[Makkai 72] Adam Makkai, *Idiom Structure in English,* Mouton, The Hague, 1972.

[Mann & Matthiessen 83] William C. Mann & Christian M.I.M. Matthiessen, *Nigel: A Systemic Grammar for Text Generation*, USC/Information Sciences Institute, Technical Report ISI/RR-83-105, Feb 1983.

[Matthiessen 84] Christian M.I.M. Matthiessen, *Systemic Grammar in Computation: the Nigel case,* USC/Information Sciences Institute, Technical Report ISI/RR-83-121, Feb 1984.

[McDonald 80] David D. McDonald, *Natural language productions as a process of decision-making under constraints*, Ph.D. thesis, Massachusetts Institute of Technology, Aug 1980.

[McDonald 83] David D. McDonald, "Natural language generation as a computational problem: an introduction," in Brady & Berwick (eds.), *Computational Problems in Discourse*, MIT Press, Cambridge, 1983.

[McDonald 85] David D. McDonald, "Description-directed natural language generation," in *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, IJCAI, 1985.

[McKeown 83] Kathleen R. McKeown, "Focus constraints on language generation," in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 582-586, IJCAI, 1983.

[McKeown 85] Kathleen R. McKeown, *Text generation: using discourse strategies and focus constraints to generate natural language text,* Cambridge University Press, Cambridge, 1985.

[Mel'chuk 81] Igor Mel'chuk, "Meaning-text models: a recent trend in Soviet linguistics," *Annual Review of Anthropology* 10, 1981, 27-62.

[Mel'chuk & Zholkovsky 84] Igor Mel'chuk & Alexander K. Zholkovsky, *Explanatory Combinatorial Dictionary of Modern Russian*, Wiener Slawistischer Almanach, Vienna, 1984.

[Mel'chuk et al. 83] Igor Mel'chuk, Lidija Iordanskaja, Nadia Arbatchewsky-Jumarie, and Adele Lessard, "Trois principes de description semantique d'une unite lexicale dans un dictionnaire explicatif et combinatoire," *Canadian Journal of Linguistics* 28, (2), 1983, 105-121.

[Patten 86] Terry Patten, *Interpreting Systemic Grammar as a Computational Representation: a problem solving approach to text generation*, Ph.D. thesis, University of Edinburgh, 1986.

[Steinacker & Buchberger 83] Ingeborg Steinacker & Ernst Buchberger, "Relating Syntax and Semantics: the syntactico-semantic lexicon of the system VIE-LANG," in *Proceedings of the First Conference of the European Chapter*, pp. 96-100, ACL, Sep 1983.

[Steinacker & Trost 83] Ingeborg Steinacker & Harald Trost, "Structural relations -- a case against case," in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, pp. 627-629, IJCAI, Aug 1983.

END

9-87

DTIC